Computer Communications xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Providing resiliency for optical grids by exploiting relocation: A dimensioning study based on ILP ☆

J. Buysse *,[1], M. De Leenheer [2], B. Dhoedt, C. Develder [2]

Ghent University – IBBT, Dept. of Information Technology (INTEC) – IBCN, Gaston Crommenlaan 8 bus 102, BE-9050 Gent, Ledeberg, Belgium

ARTICLE INFO

Article history:
Received 2 April 2010
Received in revised form 20 August 2010
Accepted 27 December 2010
Available online xxxx

Keywords:
Shared Path protection
Anycast
Grid computing
Resiliency
Relocation

ABSTRACT

Grids use a form of distributed computing to tackle complex computational and data processing problems scientists are presented with today. When designing an (optical) network supporting grids, it is essential that it can overcome single network failures, for which several protection schemes have been devised in the past. In this work, we extend the existing Shared Path protection scheme by incorporating the anycast principle typical of grids: a user typically does not care on what specific server this job gets executed and is merely interested in its timely delivery of results. Therefore, in contrast with Classical Shared Path protection (CSP), we will not necessarily provide a backup path between the source and the original destination. Instead, we allow to relocate the job to another server location if we can thus provide a backup path which comprises less wavelengths than the one CSP would suggest. We assess the bandwidth savings enabled by relocation in a quantitative dimensioning case study on an European and an American network topology, exhibiting substantial savings of the number of required wavelengths (in the order of 11–50%, depending on network topology and server locations). We also investigate how relocation affects the computational load on the execution servers. The case study is based on solving a grid network dimensioning problem: we present Integer Linear Programming (ILP) formulations for both the traditional CSP and the new resilience scheme exploiting relocation (SPR). We also outline a strategy to deal with the anycast principle: assuming we are given just the origins and intensity of job arrivals, we derive a static (source,destination)-based demand matrix. The latter is then used as input to solve the network dimensioning ILP for an optical circuit-switched WDM network.

1. Introduction

1.1. Optical grids

The very demanding requirements of several problems in domains ranging from astrophysics [1], climate modeling [2] and fluid dynamics [3] have led to the conception of grid computing. A grid consists of different heterogeneous resources (computational, storage and networking) which are geographically spread over various administrative domains, implying that resource coordination is not subject to centralized control. To interconnect the distributed resources, the optical network with Wavelength Division Multiplexing (WDM) is a suitable candidate for it, since it can support high bandwidth traffic with low latency in a reliable way. This has led to the concept of optical grids or so-called lambda grids [4,5]. While multiple alternative optical switching techniques have been proposed (including optical burst switching, OBS), in this paper we focus on circuit-switched (OCS) optical grids where wavelength connections (so-called lambdas in lambda-grids) are set-up, establishing connectivity between a source and a destination node using a two-way reservation.

One characteristic of an optical grid is the *anycast principle* which in this context means that the user is not interested in the location of the execution of his application (which we will denote as jobs), but is merely concerned with the successful execution of the jobs subject to predetermined requirements such as a fixed deadline or some other quality guarantee. To guarantee this timely delivery, we have to make sure that it is also realized in case of a resource failure (either network or computing resources). In this work we address survivability of single link failures in the optical network. There are two basic strategies to protect an optical net-

☆ The work described in this paper was partly funded by the European Commission through the 7th ICT-Framework Programme projects IST Phosphorus and BONE (Building the Future Optical Network in Europe, a Network of Excellence).

* Corresponding author. Tel.: +32 9 331 49 42; fax: +32 9 331 48 99.
E-mail addresses: jens.buysse@intec.ugent.be (J. Buysse), chris.develder@intec.ugent.be (C. Develder).

[1] J. Buysse is supported by the agency for Innovation by Science and Technology (IWT).
[2] M. De Leenheer and C. Develder are post-doctoral fellows of the Research Foundation – Flanders (FWO – Vl).

0140-3664/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2010.12.014

work, namely *restoration* and *protection* [6]. The former is a reactive procedure where connections affected by a failure are routed along an alternative path that is calculated and set up at the time of the failure. In case of protection, the backup path is pre-computed. This paper discusses two protection schemes, establishing for each primary path an associated backup path to be used whenever one of the links in the primary fails. The first protection scheme we take into consideration is the well-known scheme which we denote as Classical Shared Path (CSP) protection: wavelengths can be shared among backup paths, as long as the corresponding primary paths are link disjoint. (Its counterpart, *dedicated Path* protection, does not allow this sharing.) Our proposed second scheme, Shared Path protection with Relocation (SPR) is an extension of the CSP scheme, where instead of reserving a backup path to the end point of the primary path—being the original destination as determined by the grid scheduler—we can provide a backup path to another (possibly closer) server site, hence allowing the jobs to relocate. We quantitatively assess the benefits in terms of overall number of wavelengths used on the whole of all network links (i.e. achievable *Network Load Reduction*, NLR), as well as the potential penalty in terms of extra load on the servers receiving the relocated jobs.

To achieve these results, we show how to solve the network dimensioning problem by means of an Integer Linear Program (ILP). ILPs are presented for both Classical Shared Path protection (CSP) providing a backup path to the original end point and Shared Path protection with Relocation (SPR). Traditionally, a static demand matrix serves as input for these formulations, specifying the number of connections to set-up between each source and possible destination. However, in a grid scenario, the destination of jobs is left up to the grid scheduler (cf. anycast). Hence, we will consider a dimensioning approach starting from arrival rates specifying the job intensity per source. In Section 3 we outline a phased strategy to convert these arrival rates to a static (source,destination)-based demand matrix. Thereby, we use an ILP to find the best possible locations for the server sites. After this, we analytically compute the server capacity while meeting a predefined job loss rate. As a last step, we use simulation, assuming a certain scheduling policy, to find the resulting static demand matrix specifying the job rates exchanged between each (source,destination)-pair.

The remainder of this paper is structured as follows. First, in Section 2, we briefly discuss the possible failures which can occur in optical grids. In Section 3 we explain how to obtain a (source,destination)-based traffic matrix from a grid scenario only specifying job origins. In Section 4 we present Integer Linear Programming (ILP) formulations for dimensioning the network assuming the new SPR protection scheme, as well as the CSP benchmark case. We present an evaluation of these models by a case study in Section 5. Final conclusions are summarized in Section 6.

1.2. Related work

In [7] a survey is presented based on input of the grid community sharing their actual experience regarding fault treatment. It shows that a large part of the failures originate from hardware deficiencies (±35%), indicating the importance of our study. The relevance of the considered single link failure model is demonstrated in [8]. The authors state that in order to provide complete protection from all dual-link-failures, one may need almost thrice the spare capacity compared to a system that protects against all single-link failures. However, it has also been shown that systems designed for 100% single-link failure protection can provide reasonable protection from dual-link failures.

A large research effort has been devoted to recovery strategies resolving resource (i.e. grid server) failures. There are two strategies which aim to improve the system's performance in the presence of failure: job checkpointing and replication. Job checkpointing [9,10] periodically stores an image of the running job, which can be restored in case of a failure. In replication [11,12] a job is sent to a primary server and to a set of replication servers. In case of a failure of the primary server, its role is taken over by a replication server which continues the job execution.

In [13] several adaptive heuristics, based on both approaches and their combination were designed and evaluated. The results have shown that the overhead of periodic checkpointing can significantly be reduced when the checkpointing frequency is dynamically adapted as a function of resource stability and remaining job execution time. Furthermore, adaptive replication-based solutions can provide for even lower cost fault-tolerance in systems with low and variable load, by postponing replication according to system parameters. Finally, the advantages of both techniques are combined in the hybrid approach that can best be applied when the distributed system properties are not known in advance. Note that [13] disregards network failures, and uses a simplified network model.

In this paper, we will focus on the network aspects and consider protection against network failures (and as such is complementary to server resiliency strategies as checkpointing and replication). For a review and classification of the main optical protection techniques for the WDM-layer, we refer to [14]. We will evaluate our proposed relocation strategy SPR by formulating two ILPs. ILPs have been widely exploited in previous works to find a optimal solution to a certain network design and planning problem. The main advantages of these kind of formulations is the easy way of adapting the description of the network environment: cost functions, wavelength conversion, protection scheme, etc.

These ILP formulations can be divided into two categories: Flow Formulation (FF) and Route Formulation (RF). The authors of [15] have investigated these formulations in unprotected networks to conclude that although they have the same computational complexity, RF has the advantage of reducing the number of variables by imposing a restriction on the number of allowable paths between a source and a destination. In [16] the authors focus on the computational efficiency of the ILP model in order to provide a more effective tool for planning. The formulation exploits flow aggregation and consists in a new ILP formulation that can reach optimal solutions with less computational effort compared to other ILP approaches. Yet, the solution of the so-called source formulation ILP in [16] requires a post-processing step to find the actual routing and wavelength assignment (RWA) and it does not consider resilient network dimensioning.

In this paper, we stick to the traditional source–destination method based on flow formulation, where the CSP case is largely based on the ILP presented in [17]. There the authors investigate the problem of fault management in a meshed WDM network with failures due to fiber cuts: both ILP and heuristic solutions are examined and their performance is compared through numerical examples.

Note that the current paper is an extended version of [18] where we presented some preliminary results on the subject. Since then, we have updated the ILP formulation for both the CSP and SPR cases, which is now more compact and reduces the required number of variables. We here also provide a significantly more extensive result set: we discuss the influence of providing more server sites, the extra load needed on the server sites, influence of topology structure, etc.

2. Failures in optical grids

Network failures in optical networks are either known in advance (planned failures) and some preventive measures can be

Please cite this article in press as: J. Buysse et al., Providing resiliency for optical grids by exploiting relocation: A dimensioning study based on ILP, Comput. Commun. (2011), doi:10.1016/j.comcom.2010.12.014

taken to overcome them, or they are unplanned and caused by erratic events such as natural disasters and fiber cuts. From a network provider's point of view, it is impossible to devise preplanned protection schemes for all imaginable network failures, and hence the most occurring failures are split up into various restricted failure scenarios to be overcome in a gracious manner. For network resources, typically cable cuts and equipment failures are the most frequent and two scenarios are considered:

1. *Single link failure*: a link between two adjacent network nodes fails and consequently no information can be sent between them. Schemes protecting against these kind of failures can reroute around the end nodes of the failed link (Fig. 1(a)) or find a new path from the source to destination (Fig. 1(b)).
2. *Single node failure*: a network element fails and hence all its incident links are out of service (Fig. 1(c)).

The aforementioned protection scheme, Classical Shared Path protection (CSP), is subdivided in the first failure class as is our newly proposed scheme, Shared Path protection with Relocation (SPR), for the very reason that it is an extension of CSP. We denote a primary path as the path which is used in the failure free scenario and it corresponding backup path as the path which is used when a single-link failure occurs on that primary path. As indicated before, we are dealing with a Shared path protection scheme which indicates that two primary paths P_1 and P_2 can be protected by two partially overlapping backup paths B_1 and B_2 as long as P_1 and P_2 are link disjoint.

$$R_1 \cap R_2 \neq \emptyset \Rightarrow P_1 \cap P_2 = \emptyset \quad (1)$$

2.1. Shared Path protection with relocation

In the CSP scheme, a primary path and its corresponding backup path end at the same node (in this case some grid server site) and two backup paths can share wavelengths as long as their corresponding primary paths are link disjoint. We will relax the first constraint so that the endpoints of a primary and backup path can end in different grid server sites, as to potentially reduce the network load. This implies relocation of grid jobs from the primary server site to an alternate site for which we could create a backup path comprising fewer hops (not including any of the primary links) or finding a backup path where more wavelengths can be shared (i.e. inferring no additional cost because they are already installed for another backup path). This relocation is possible by the grid specific anycast principle: when a user creates a job, several resources are able to execute it and only one of them is chosen, generally by the grid scheduler. Hence, as illustrated in Fig. 2, in case of a network failure on the primary path we could relocate

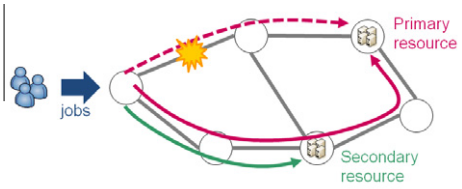


Fig. 2. In a Classical Shared Path protection scheme (CSP) a primary path is protected by a link disjoint backup path. By allowing the backup path to end in a server different from the primary server, we can achieve a network load reduction. This resilience scheme is called Shared Path protection with Relocation (SPR).

the job to another computing resource. Still, this could cause a trade-off between lowering network resources (fewer wavelengths) and potentially increasing resource capacity: we have to cater for extra computing power at the relocation server to process relocated jobs. Note however that such additional server capacity will be required anyhow to cope with grid resource failures.

3. Deriving a (source,destination)-traffic matrix from anycast grid traffic

Our goal is to evaluate the above-mentioned relocation scheme against Classical Shared Path protection, from a network dimensioning perspective. Hence, we will employ ILP formulations to derive the required amount of wavelengths needed to equip for a given connection demand between (source,destination)-pairs. However, in an optical grid scenario where the anycast principle applies, the traffic is rather specified by the number of jobs arriving at given source sites and the destination can essentially be freely chosen among server sites. Hence, we need to convert this anycast traffic specification to a clearly defined (source,destination)-based traffic matrix as required for network dimensioning algorithms (such as ILP). We now will present a methodology realizing this conversion, before discussing the network dimensioning in Section 4.

To obtain our traffic matrix, we resorted to an iterative approach. This is discussed in detail in [19], and summarized below. The subsequent phases followed stem from the realization that three aspects are important when trying to obtain a (source,destination)-based traffic matrix from the demand vector:

1. The location of the grid server sites, which are capable of executing the jobs.
2. The amount of servers at each of the chosen server sites.
3. The scheduling algorithm: the policy the grid management enforces to distribute the jobs among the different server sites.

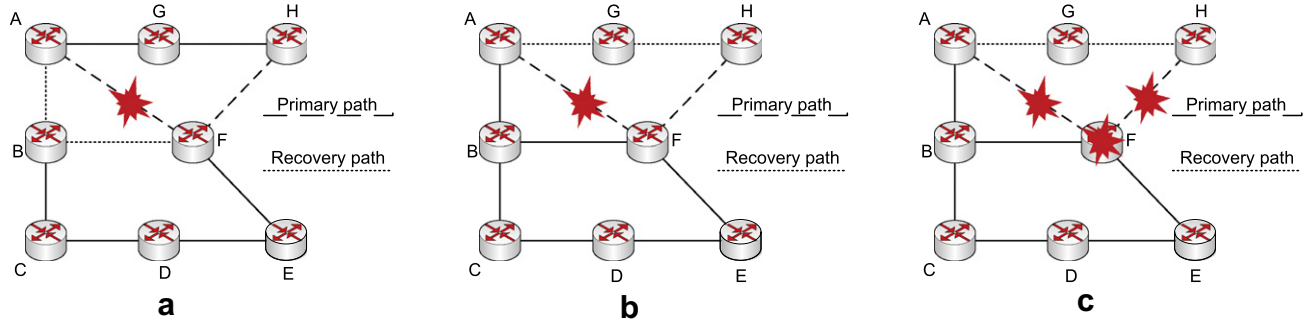


Fig. 1. Failure scenarios and recovery paths in a communication network for a connection from A to H. (a) *Single link failure with link protection*: When the link A–F fails, this link is bypassed by the links A–B and B–F after which the original path is reused. (b) *Single link failure with path protection*: When a link on the path from A to H fails, the backup path A–G–H is taken. (c) Single node failure, causing two links to fail. When node F fails, the recovery path A–G–H is taken.

Please cite this article in press as: J. Buysse et al., Providing resiliency for optical grids by exploiting relocation: A dimensioning study based on ILP, Comput. Commun. (2011), doi:10.1016/j.comcom.2010.12.014

Thus, the first steps are to decide where to locate the server sites and how many server CPU's to install at each site (e.g. while meeting a maximum job loss rate criterion).

3.1. Find the K best server locations

Choosing the optimal choice for the server locations is a *K*-medoid problem: the goal is to find *K* clusters, where the nodes in each cluster are grouped together according to a specified metric and where the cluster centers represent the chosen server sites. We have formulated this as a compact ILP shown below, making the simplifying assumption that site *i* sends all its jobs to the same server (which may not be the case in reality, depending on the scheduling policy, described in Section 3.3).

The decision variables deciding on the server site locations are:

- $T_j = 1$ if and only if site j is chosen as a server site location, else 0.
- $S_{i,j} = 1$ if and only if site j is the target server for traffic from site i, else 0.

The given input parameters to base these decisions on are:

- λ_i is the job arrival rate at site j ($i = 1 \ldots N$).
- $H_{i,j}$ is the routing distance (typically hop count) from site i to site j ($i,j = 1 \ldots N$).
- K is the number of server sites to choose.

The objective function of the ILP is given in Eq. (2), the constraints are in Eqs. (3)–(5).

$$min\left(\sum_i \sum_j \lambda_i \cdot H_{i,j} \cdot S_{i,j}\right) \tag{2}$$

$$\sum_j T_j = K \tag{3}$$

$$\sum_j S_{i,j} = 1 \quad \forall i \tag{4}$$

$$S_{i,j} \leqslant T_j \quad \forall i,j \tag{5}$$

3.2. Determining the server capacities

We continue with dimensioning the processing power at each server site, i.e. the number of CPUs. We have made some assumptions which appear to be realistic [20]: we assume Poisson arrivals and exponentially distributed service times. With these assumptions we solve the well-known ErlangB formula (6) to establish the total number n of servers needed to meet a maximum job loss rate of x%. We subsequently distribute that amount of n CPUs among the server sites, proportionally to the cluster arrival rate at each server site (thus installing the most CPUs where the most traffic is arriving, as [19] indeed showed this choice results in lower network loads).

$$ErlangB(\lambda, \mu, n) = \frac{\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}}{\sum_{k=0}^{n} \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!}} = x \tag{6}$$

3.3. Scheduling policy

We have adopted a *mostfree* scheduling policy (see [19]): first try the server nearest (in terms of hop count, hence denoted as 'local' server site) to the job's originating site. If this 'local' server site is not available, then choose a free CPU at server site f, where f is the server site with the highest number of free server CPUs, in an attempt to avoid overloading sites and thus limiting non-local job execution. In this step we have resorted to simulations because of the anycast principle: it is hard to obtain accurate estimates for the inter-site traffic using analytical techniques (although that under certain assumptions, numerical calculation can be achieved [21]). Note that this scheduling policy holds at runtime and so the assumption that each source site sends to the same server made in Section 3.1 does not necessarily hold. Yet, if the number of servers is appropriately chosen, the majority of the jobs should end up being executed at the closest server (see [19]).

After this step we know how many jobs are exchanged between every grid node pair in the considered network. By appropriately scaling with the job data sizes and rounding these numbers, we finally end up with a demand matrix containing a number of connections between each grid node pair.

4. Network dimensioning model

We investigate a network design model with a static traffic matrix in which a known set of connection requests is assigned to the network. Each connection represents a point-to-point light path (circuit) from a source to a destination, able to transport a given capacity. Furthermore, we assume in this paper a so-called virtual wavelength path (VWP) network [17], implying that all optical cross-connects (OXC) are able to perform wavelength conversion. Note that if OXCs do not support wavelength conversion, the wavelength continuity constraint must hold and the resulting network is a plain wavelength path (WP) network.

Our topology is modeled as a graph $G = (V,E)$ where the links are represented by a directed edge $(i,j) \in E$ (with $|E| = L$), while the vertices $v \in V$ (with $|V| = N$) represent the OXCs. The static traffic matrix is converted into a list of connection objects $\beta = \{\phi_1, \phi_2, \ldots, \phi_n\}$ where a connection ϕ_c corresponds a unit demand requiring a single wavelength path, identified by its index c. Two connections can have the same source and the same destination.

We define the following variables:

- $p_{i,j}^{\phi}$: binary decision variable which is 1 if link (i,j) is used for the primary path for connection ϕ.
- $r_{(i,j)}^{\phi}$: binary decision variable which is 1 if link (i,j) is used as part of a protection path for connection ϕ.
- m_j^{ϕ}: binary decision variable which is 1 if node j is a backup resource which is used for connection ϕ.
- $\pi_{i,j}$: integer auxiliary variable, the total number of wavelengths on link (i,j) used for a backup path.
- $P_{i,j}$: integer auxiliary variable, the total number of wavelengths on link (i,j) used for a primary path.
- $\Theta_{(i,j),(k,l)}^{\phi}$ is an integer variable introduced to calculate the number of shared wavelengths.

4.1. ILP formulation

The objective function (7) expresses that we want to minimize the total number of primary and backup wavelengths:

$$\min\left(\sum_{i,j} \pi_{i,j} + \sum_{i,j} P_{i,j}\right) \tag{7}$$

Constraints (8) express the demand constraints and flow conservations for the primary paths. When j is the source node of connection ϕ ($j = s$) then we should only have a flow originating from that source. If j is the destination of ϕ ($j = d$) then this node should be the ending node of the flow. In the last case, where the j is an OXC, any connection arriving should also leave again. Similarly,

Please cite this article in press as: J. Buysse et al., Providing resiliency for optical grids by exploiting relocation: A dimensioning study based on ILP, Comput. Commun. (2011), doi:10.1016/j.comcom.2010.12.014

the constraints (9) are the flow conservations for the backup paths, where the m_j^ϕ variable will decide which node is the destination and will depend on whether we are considering CSP or SPR.

$$\sum_{i:(i,j)\in E} p^\phi_{(i,j)} - \sum_{k:(j,k)\in E} p^\phi_{(j,k)} = \begin{cases} -1: & j = s \\ +1: & j = d \\ 0: & \text{else} \end{cases} \quad \forall \phi \in \beta, \forall j \in V \tag{8}$$

$$\sum_{i:(i,j)\in E} r^\phi_{(i,j)} - \sum_{p:(j,p)\in E} r^\phi_{(j,p)} = \begin{cases} -1: & j = s \\ m_j^\phi: & \text{else} \end{cases} \quad \forall \phi \in \beta, \forall j \in V \tag{9}$$

The next constraints (10) express that a primary path and a backup path cannot overlap.

$$r^\phi_{(i,j)} + p^\phi_{(i,j)} \leqslant 1 \quad \forall \phi \in \beta, \forall (i,j) \in E \tag{10}$$

In (11) we introduce the binary variable $\Theta^\phi_{(i,j),(k,l)}$ which is 1 if and only if for connection ϕ link (k,l) is protected by link (i,j). These variables are used in (12) to bound the $\pi_{(i,j)}$ variables which count the shared backup wavelengths for a link (i,j).

$$\Theta^\phi_{(i,j),(k,l)} + 1 \geqslant r^\phi_{(i,j)} + p^\phi_{(k,l)} \quad \forall \phi \in \beta, \forall (i,j),(k,l) \in E \tag{11}$$

$$\pi_{(i,j)} \geqslant \sum_\phi \Theta^\phi_{(i,j)(k,l)} \quad \forall (k,l) \in E, \forall (i,j) \neq (k,l) \in E \tag{12}$$

In the case of CSP we enforce that the primary server and backup server need to be the same by Eq. (13).

$$m_j^\phi = \begin{cases} 1 & \text{if } j \text{ is the primary server of } \phi \\ 0 & \text{else} \end{cases} \tag{13}$$

On the other hand, to achieve SPR we replace (13) with (14), (15) to let the ILP freely decide which backup server to use.

$$\sum_{\delta\in\Delta} m_\delta^\phi = 1, \quad \forall \phi \in \beta \tag{14}$$

$$m_\delta^\phi = 0, \quad \forall \delta \notin \Delta \tag{15}$$

4.2. Complexity

According to [17], the complexity of an ILP heavily depends on the number of variables and to a lesser extent on the number of constraints. The number of variables of the ILP formulations are the same for both the CSP and SPR cases, while only the number of constraints differ. Nevertheless, there is a big difference in the running time of the CSP vs. the SPR: running a CSP instance with the same input parameters takes much longer than an instance of SPR.

The number of variables is

$$2|E| \times (|\beta| + 1) + |\beta| \times (|V| + |E|^2)$$

and depends mostly on the number of desired connections and the topology. The number of constrains for CSP is

$$|\beta| \times (2|V| + |E|) + |E|^2 \times (|\beta| + 1)$$

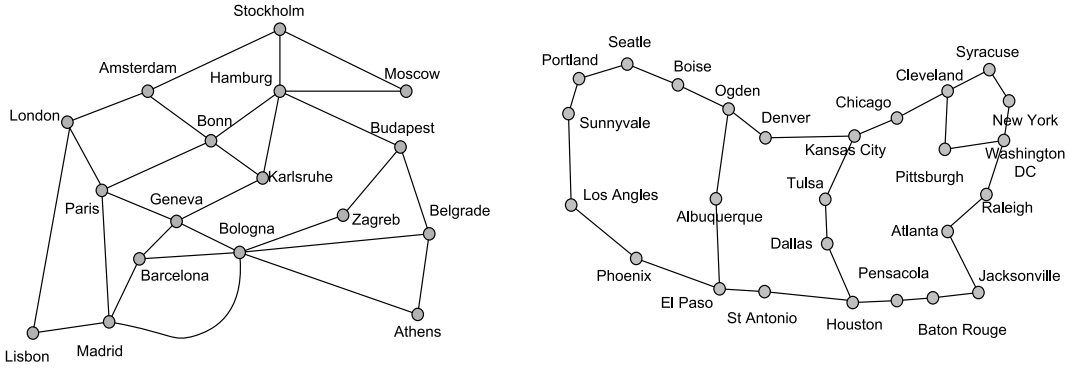
If we want to achieve SPR we have add $|\beta| + |\Delta|$ more constraints. We notice that the ILP is not very scalable (quadratic in the number of links) and will not suffice to deal with larger instances. Future work could include investigating a scalable heuristic or an inquiry on how to convert this ILP into a more scalable formulation (e.g. column generation).

5. Case study

We have considered the two topologies depicted in Fig. 3, where each link is supposed to be bidirectional. Fig. 3(a) is based on the Géant 2 network topology and its associated various national research and education networks (NRENs) and consists of 17 nodes and 56 links. Fig. 3(b) is based on the National Lambda Rail (NLR) which provides a testbed for advanced research at over 280 universities, US government laboratories and advanced programs across the United States and consists of 27 nodes and 60 links. For each topology, we have generated 10 random arrival rate files, containing for every possible source site the rate of jobs it needs to send out. By applying the strategy explained in Section 3 we end up with 10 different demand matrices (with increasing number of connections) for each network with respectively 3, 5 and 7 server sites. These static demand matrices served as input for the ILP and their results are presented in the sections below. (Note that for a given number of unit connection demands we chose not to present average results over multiple random instances, since the chosen server sites may differ among them.) We will use the notation N_x^y as a network with x server sites and a connection demand of y connections.

5.1. Influence of relocation

We first discuss the results obtained for the EGEE Géant-based network. In Fig. 4 we plot the total number of wavelengths, summed over all links, being used for all the primary and backup



a EGEE topology **b** USNLR topology

Fig. 3. Topologies for the case studies. The first is based on the EGEE GEANT network consisting of 17 nodes and 56 links. The second is the US National Lambda Rail (USNLR) consisting of 27 nodes and 60 links.

Please cite this article in press as: J. Buysse et al., Providing resiliency for optical grids by exploiting relocation: A dimensioning study based on ILP, Comput. Commun. (2011), doi:10.1016/j.comcom.2010.12.014

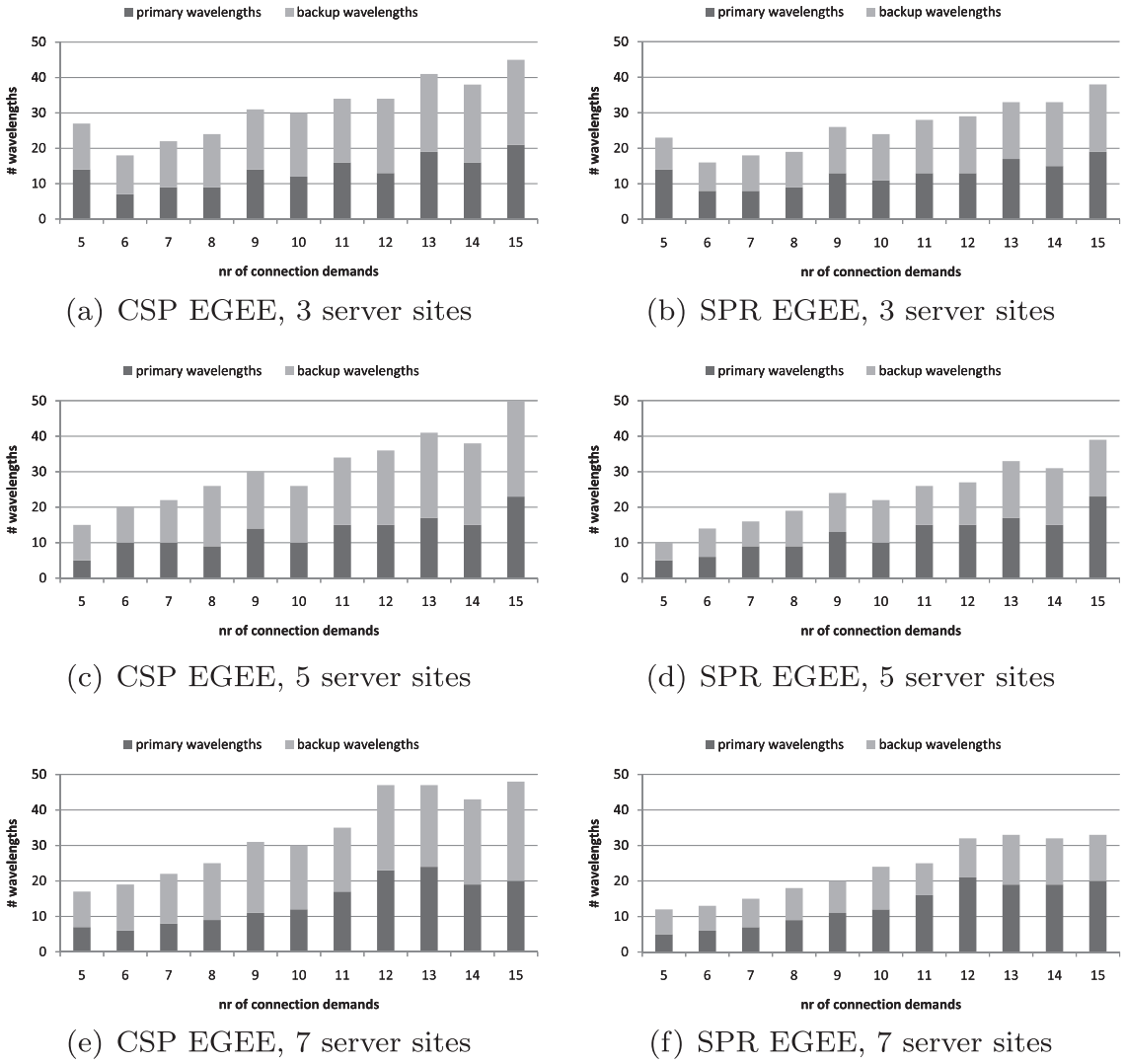


Fig. 4. The total number of wavelengths for both the CSP and SPR case, for the EGEE network with 3, 5 or 7 server sites. Although there is little or no difference in the amount of primary wavelengths between both CSP and SPR, the number of backup wavelengths for SPR amounts on average to 24% of the number of backup wavelengths of CSP, with peak up to 50%. Note that each bar is the result of a single dimensioning outcome, hence the non-monotonic increase for increasing number of unit connection demands.

paths. As expected, with an increasing load, the number of required network resources tend to grow. (Note that the increase is not monotonic, given that we are considering single random cases: thus it is possible that comparing two cases with different number of unit connection demands, the one with the higher demand not necessarily requires more wavelengths.)

Comparing the amount of primary wavelengths used in CSP with the amount of primary wavelengths in SPR we see that there is little or no difference and this observation is independent on the number of server sites which have been chosen. This means it does not often happen that SPR finds a primary path (different from the CSP case) to create more opportunities for sharing wavelengths among different connections' backup paths.

Yet, the number of backup wavelengths can be drastically decreased by employing relocation:

- For N_3^y an average decrease of 24% with a peak of 33%.
- For N_5^y an average decrease of 36% with a peak of 50%.
- For N_7^y an average decrease of 44% with a peak of 55%.

There are two possible reasons (which may apply simultaneously) why relocating to another site consumes fewer backup wavelengths:

1. *Closer backup site:* Relocating a job allows to establish a backup path to a another (backup) server site which—considering a failure of any of the primary path's links—is closer in terms of hop count (and thus fewer wavelengths summed over all links), e.g. a server that lies on CSP's backup path to the primary server.
2. *More sharing:* A connection ϕ's path to a server site, other than the primary one, could contain many backup wavelengths for connections having a primary paths disjoint from ϕ's. Hence, a larger portion of such a backup path may comprise wavelengths shared with others, requiring fewer wavelengths to be set-up exclusively for ϕ.

As can be noted, increasing the server sites has an positive influence on the reduction of backup wavelengths. We will come back to this in Section 5.2.

Looking at Fig. 5 for the USNLR network, and comparing with the EGEE results, we observe substantial difference between the absolute numbers of wavelengths between the EGEE and the USNLR cases. This obviously stems from the highly different network topologies: the EGEE topology is more meshed while the USNLR topology is much sparser, resembling a composition of rings. Hence the cycle formed by a primary and its corresponding backup path covers ring-like structures which comprise consider-

Please cite this article in press as: J. Buysse et al., Providing resiliency for optical grids by exploiting relocation: A dimensioning study based on ILP, Comput. Commun. (2011), doi:10.1016/j.comcom.2010.12.014

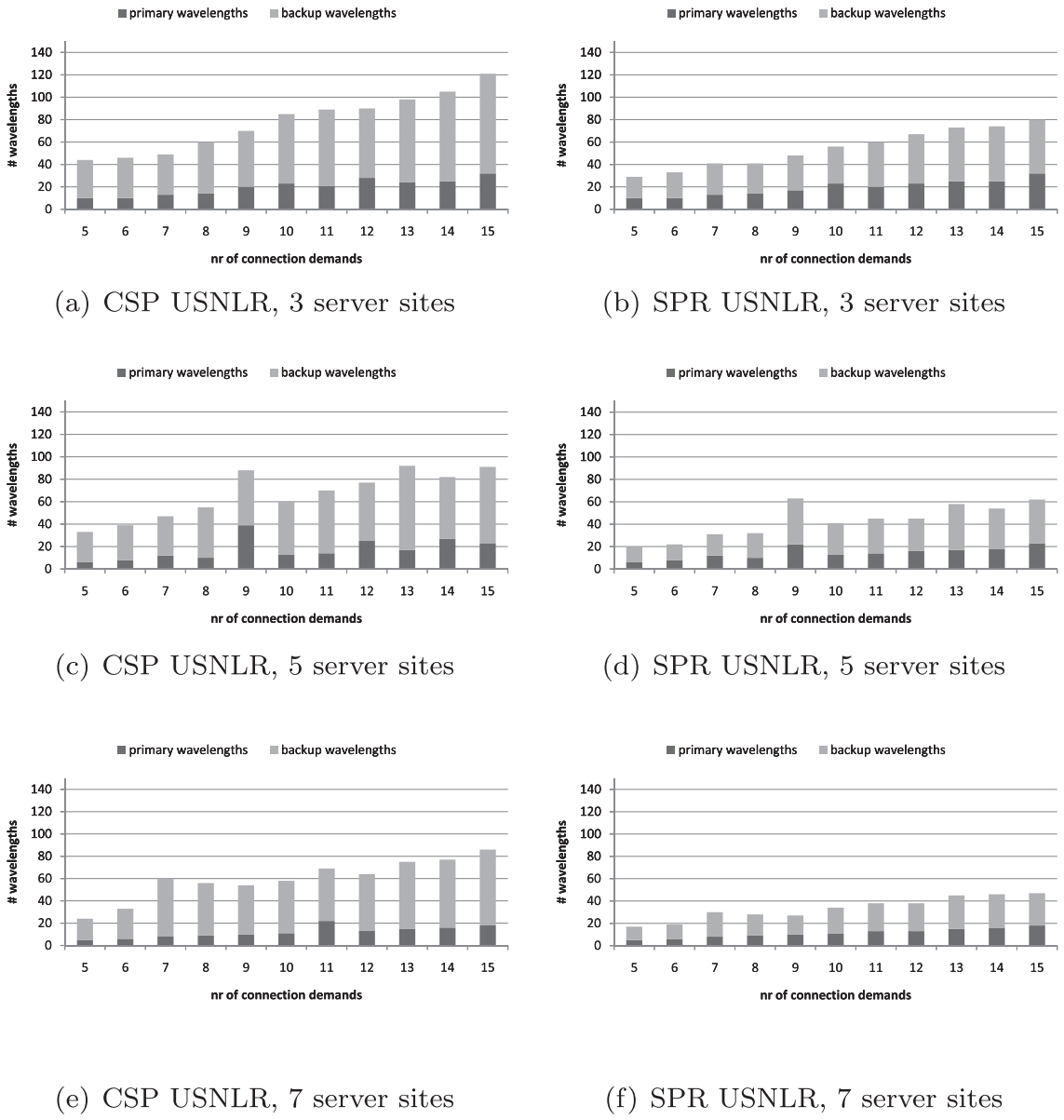


Fig. 5. The total number of wavelengths for both the CSP and SPR case, for the USNLR network with 3, 5 or 7 server sites. Similar observations apply as for the EGEE network.

ably more hops than in a highly meshed topology. Apart from the relatively higher number of backup wavelengths, similar observations as for the EGEE network can be made:

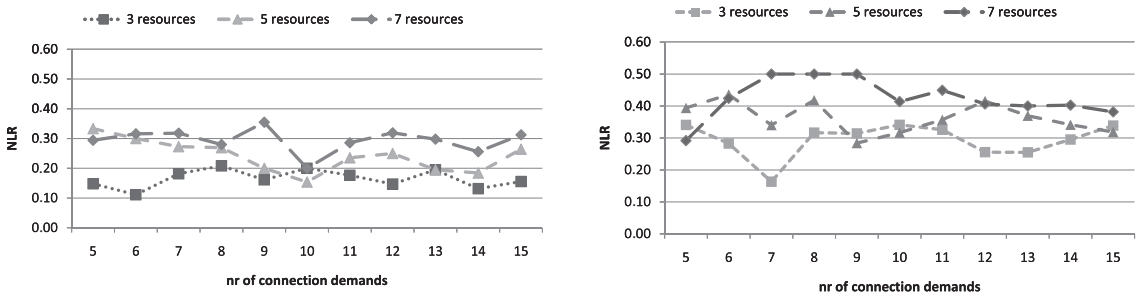
- With an increasing load, we generally achieve a higher number of required wavelengths.
- Comparing CSP with SPR, we see that we can drastically reduce the number of wavelengths.
  - This decrease is not induced by a decrease of primary wavelengths, because that number stays the same in most cases for CSP and SPR.
  - The decrease mainly stems from a reduction in backup wavelengths (for USNLR up to 61%) by either relocating to another closer server site or exploiting a sharing possibility which was not possible in the CSP case.

5.2. Network load reduction

As pointed out in Section 5.1, relocation achieves a lower number of consumed wavelengths—mainly induced by the decrease in backup rather than primary wavelengths—which we express formally as network load reduction (NLR) in Eq. (16). We have plotted this NLR for both the EGEE and USNLR network, in Fig. 6(a) and (b) respectively, for N_x^y, $x \in \{3,5,7\}, y \in [5,15]$. We note that it seems that when employing more servers, the NLR increases. A reason for this may be that using more servers implies a higher probability of encountering another server on the backup path to the original one, and thus relocation is favorable. (Nevertheless, in some rare cases, having fewer servers does amount to a higher NLR; which may be due to single random demand creation, and the fact that having different server locations will amount to a different traffic matrix instance, cf. scaling to conform to integer demands.).

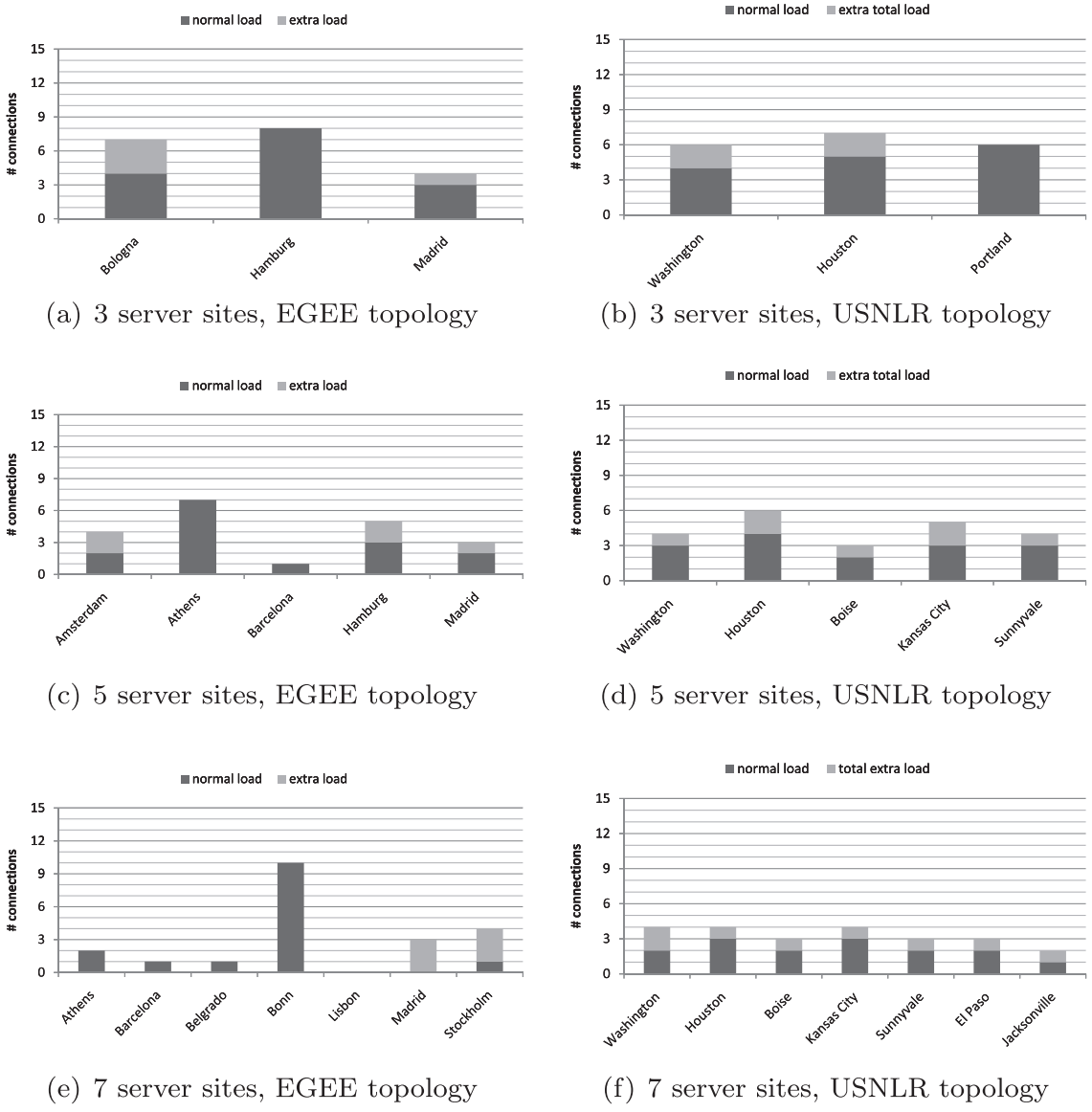
Considering the results for the USNLR network in Fig. 6(b), we note that qualitatively, the same observations apply as in the EGEE case. Yet, when comparing the NLR for a N_x^y case for both networks we notice that the USNLR more often than not has a larger NLR (up to 50%). The reason for this can be found in the topology structure. The EGEE is a more meshed topology (higher average node degree: 3.29 for EGEE vs. 2.22 for USNLR). Therefore in the USNLR case, the cycle composed by a primary path and its corresponding backup path will be quite long on average. Consequently, providing a back-

Please cite this article in press as: J. Buysse et al., Providing resiliency for optical grids by exploiting relocation: A dimensioning study based on ILP, Comput. Commun. (2011), doi:10.1016/j.comcom.2010.12.014



(a) EGEE topology (b) USNLR topology

Fig. 6. The Network Load Reduction (NLR) achieved by relocation for both the topologies. By employing more servers sites we can achieve a higher NLR: for 7 server sites the savings achieved by relocation (SPR) compared to Classical Shared protection (CSP) are more substantial than for 5 or 3 server sites. Comparing both networks, we observe that in general we can achieve a higher NLR in the sparser USNLR topology.



(a) 3 server sites, EGEE topology (b) 3 server sites, USNLR topology

(c) 5 server sites, EGEE topology (d) 5 server sites, USNLR topology

(e) 7 server sites, EGEE topology (f) 7 server sites, USNLR topology

Fig. 7. The server site load in terms of number of arriving connections comprises: (i) in black the number of connections it receives in the failure free case, (ii) in the gray the maximum number of extra connections due to a single link fault. Considering on the one hand the EGEE topology cases, when putting three servers sites into service (Fig. 7), we note that Bologna receives 20% and Madrid 7% of the total server load as extra load, while Hamburg does not need to cater for anything except its failure free load. Increasing the server count to 5 decreases this average extra extra load from 9% to 7% and the peak extra load of 20% is gone (the peak now is only 13%). The case introducing 7 server locations (Fig. 7) exhibits a dedicated relocation server: this server is solely used to cope with relocated jobs. For the USNLR topologies on the other hand, we see that increasing the server site count, levels the failure free demand per server. The extra load induced by relocation averages to 8% per server and never exceeds 13% of the total requested connections. Also, we note that there are no servers exclusively used for relocation.

Please cite this article in press as: J. Buysse et al., Providing resiliency for optical grids by exploiting relocation: A dimensioning study based on ILP, Comput. Commun. (2011), doi:10.1016/j.comcom.2010.12.014

up path to another resource can drastically reduce the number of links necessary to that closer backup resource, especially when that new backup resources lies on the original backup path. That is how the the reduction in backup wavelengths demonstrated in Section 5.1 can be explained.

$$NLR = 1 - \frac{\text{total number of wavelengths SPR}}{\text{total number of wavelengths CSP}} \quad (16)$$

5.3. Extra server capacity

As previously demonstrated, by relocating to another server site instead of the one originally (i.e. under failure free condition) proposed by the grid scheduler, a significant reduction in network resources can be achieved. But there is a trade-off: the relocation server receives more jobs than originally intended and thus, needs to reserve some spare capacity in order to execute the relocated jobs. Fig. 7 shows for the EGEE topology the maximum amount of connections a server site receives for the cases with three (Fig. 7(a)), five (Fig. 7(c)) or seven server sites (Fig. 7(e)) for the demand case of 15 unit connections. The black part is the load in failure free conditions, the grey part is the maximum of extra load it receives due to a single link fault.

For N_3^{15} (Fig. 7(a)) we see that every server site has a failure free load and two sites have an extra load. For Bologna and Madrid this extra load is respectively 3/4 and 1/3 times its failure free load. Actually 1 connection is only 1/15 of the total load and if we would express each extra load relative to the load over all servers we end up that every relocation server only caters for respectively 20% and 7% of the total load, while Hamburg does not need to cater for any excess load.

Looking at the N_5^{15} case (Fig. 7(c)), we see that the load gets more evenly distributed over the different server sites, as is also the case with the extra relocation load (where the average extra load amounts to only 7% of the toal load).

The last case is N_7^{15} (Fig. 7(e)). We notice that not every server receives a failure free load which can be attributed to the high node degree of the network and the small number of source nodes of the network: adding an extra server site to the topology, e.g. going from a N_x^y to a N_{x+1}^y, does not affect the already established clusters of the N_x^y topology. Adding an extra cluster does not mean that a large enough portion of the source nodes is now closer to that extra server site. As a consequence, in the step where the server capacities are chosen (step 3.2), the extra cluster does not have a large enough cluster arrival rate and hence, the installed server capacity will be negligible compared to the installed server capacities of the other cluster. Consequently, in the scheduling step (where the *mostfree* algorithm is used), only a small fraction of the jobs will be sent to this extra server site (given the integer nature of our connection demand matrix, the rounding process will lead to 0 unit connections sent to that server). This is also the reason why Bonn receives a large failure free load: it is the site where the most capacity is installed (in the server dimensioning step). However we do see that a server site can be used as dedicated relocation server site (cf. Madrid) which only receives load in a link-failure scenario.

Focusing on the server site loads for the USNLR case (Fig. 7(b), (d), (f)), we see they are somewhat different in nature compared to the EGEE case. For all three cases, each server site receives jobs. The discussion above (for EGEE) does no longer apply for this much sparser (ring-like) USNLR topology. It is clear that adding an extra server site, e.g. going from a N_x^y to a N_{x+1}^y, is far more profitable in this sparse network case and attracts a reasonably large arrival rate. Therefore the scheduling and rounding steps of the iterative algorithm in Section 3 do not result into zero unit connection demands towards these added server sites. Accordingly, the notion of an exclusive relocation site disappears. Also, every resource site receives almost an equal part of the relocated jobs (except the N_3^{15} case where Portland does not receive this extra load). Every extra load is either 1 or 2 extra connections which caters for only 6% and 13% of the total requested connections between source and destination sites.

6. Conclusion

In this work we have described an alternative method for path protection against single link failures in an optical grid scenario. Whereas traditional protection schemes try to reserve backup capacity to the original destination of the primary path, we have accounted for the grid-specific anycast principle (stating that there are several destinations possible for a job to be executed). Therefore, in case of a network failure, we allow to relocate the job to an alternative server site, and as such are able to reduce the bandwidth (wavelengths) to be allocated for the backup path. We have described ILP models for both the traditional shared protection scheme, as well as shared protection with relocation. Our case study pointed out that on average we can achieve a reduction of the total number of necessary wavelengths (network load reduction, NLR) in the range of 11% to 50%, depending on the amount of server sites that have been chosen and the network topology (with a higher NLR for a sparser topology). A sparse network can benefit more of relocation due to the fact that it is more likely to encounter another server on the backup path found in the CSP case.

The NLR is caused by the reduction of backup wavelengths, rather than primary wavelengths. However, the relocation strategy requires adjusted capacities of the relocation servers, since they have to be able to handle these relocated jobs. The amount of extra load is dependent on the number of server sites which have been chosen and again the topology structure. On the one hand, for a meshed European network, we perceived that when selecting 3 server sites we need to provide up to about 20% of the total load as extra capacity). When we increased the number of server sites to 5, this maximum extra load decreased down to 13%. Increasing the number of server sites more is not beneficial for the server resource utilization anymore, while it is for the network dimensions.

On the other hand, for a sparser US network case study increasing the server site count rather evenly distributes the (failure free) load over the various server sites, as well the extra relocation load. This extra server load now amounts to between 6% and 13%.

References

[1] G. Allen, G. Daues, J. Novotny, J. Shalf, The astrophysics simulation collaboratory portal: A science portal enabling community software development, in: Proc. 10th IEEE Int. Symp. High Performance Distributed Computing (HPDC 2001), IEEE Computer Society, Washington, DC, USA, 2001, p. 207.

[2] B. Allcock, I. Foster, V. Nefedova, A. Chervenak, E. Deelman, C. Kesselman, J. Lee, A. Sim, A. Shoshani, B. Drach, D. Williams, High-performance remote access to climate simulation data: a challenge problem for data grid technologies, in: Proc. ACM/IEEE Conf. Supercomputing (SC 2001), Denver, CO, USA, 2001, pp. 46–46, doi:10.1145/582034.582080.

[3] S. Barnard, R. Biswas, S. Saini, R. Van der Wijngaart, M. Yarrow, L. Zechtzer, I. Foster, O. Larsson, Large-scale distributed computational fluid dynamics on the information power grid using globus, in: Proc. 7th Symp. Frontiers of Massively Parallel Computation (FRONTIERS 1999), Washington, DC, USA, 1999, p. 60.

[4] M. De Leenheer, C. Develder, T. Stevens, B. Dhoedt, M. Pickavet, P. Demeester, Design and control of optical grid networks (invited), in: Proc. 4th Int. Conf. on Broadband Networks (Broadnets 2007), Raleigh, NC, 2007, pp. 107–115, doi:10.1109/BROADNETS.2007.4550413.

[5] D. Simeonidou, R. Nejabati, G. Zervas, D. Klonidis, A. Tzanakaki, M.J. O'Mahony, Dynamic optical-network architectures and technologies for existing and emerging grid services, IEEE J. Lightwave Technol. 23 (10) (2005) 3347–3357, doi:10.1109/JLT.2005.856254.

Please cite this article in press as: J. Buysse et al., Providing resiliency for optical grids by exploiting relocation: A dimensioning study based on ILP, Comput. Commun. (2011), doi:10.1016/j.comcom.2010.12.014

[6] D. Colle, S. De Maesschalck, C. Develder, P. Van Heuven, A. Groebbens, J. Cheyns, I. Lievens, M. Pickavet, P. Lagasse, P. Demeester, Data-centric optical networks and their survivability, IEEE J. Sel. Areas Commun. 20 (1) (2002) 6–20, doi:10.1109/49.974658.

[7] R. Medeiros, W. Cirne, F. Brasileiro, J. Sauvé, Faults in grids: Why are they so bad and what can be done about it? in: Proc. 4th Int. Workshop on Grid Computing (GRID 2003), 2003, p. 18.

[8] M. Sivakumar, C. Maciocco, M. Mishra, K.M. Sivalingam, A hybrid protection-restoration mechanism for enhancing dual-failure restorability in optical mesh-restorable networks, in: Proc. Optical Networking and Commun. (OptiComm 2003), Dallas, TX, USA, 2003, pp. 37–48, doi:10.1117/12.533166.

[9] S. Agarwal, R. Garg, M.S. Gupta, J.E. Moreira, Adaptive incremental checkpointing for massively parallel systems, in: Proc. 18th ACM Int. Conf. Supercomputing (ICS 2004), Malo, France, 2004, pp. 277–286, doi:10.1145/1006209.1006248.

[10] J.W. Young, A first order approximation to the optimum checkpoint interval, Commun. ACM 17 (9) (1974) 530–531, doi:10.1145/361147.361115.

[11] Y. Li, M. Mascagni, Improving performance via computational replication on a large-scale computational grid, in: Proc. 3rd IEEE/ACM Int. Symp. Cluster Computing and the Grid (CCGrid 2003), 2003, pp. 442–448, doi:10.1109/CCGRID.2003.1199399.

[12] C.-J. Hou, K.G. Shin, Replication and allocation of task modules in distributed real-time systems, in: Proc. 24th Int. Symp. on Fault-Tolerant Computing (FTCS-24), 1994, pp. 26–35, doi:10.1109/FTCS.1994.315660.

[13] M. Chtepen, F.H.A. Claeys, B. Dhoedt, F. De Turck, P. Demeester, P.A. Vanrolleghem, Adaptive task checkpointing and replication: Toward efficient fault-tolerant grids, IEEE Trans. Parallel Distri. Syst. 20 (2) (2009) 180–190, doi:10.1109/TPDS.2008.93.

[14] G. Maier, A. Pattavina, S. De Patre, M. Martinelli, Optical network survivability: Protection techniques in the WDM layer, Photonic Network Commun. 4 (3–4) (2002) 251–269, doi:10.1023/A:1016047527226.

[15] N. Wauters, P. Demeester, Design of the optical path layer in multiwavelength cross-connected networks, IEEE J. Sel. Areas Commun. 14 (5) (1996) 881–892, doi:10.1109/49.510911.

[16] M. Tornatore, G. Maier, A. Pattavina, WDM network design by ILP models based on flow aggregation, IEEE/ACM Trans. Network 15 (3) (2007) 709–720, doi:10.1109/TNET.2007.893158.

[17] H. Zang, C. Ou, B. Mukherjee, Path-protection routing and wavelength assignment RWA in WDM mesh networks under duct-layer constraints, IEEE/ACM Trans. Network 11 (2) (2003) 248–258, doi:10.1109/TNET.2003.810313.

[18] J. Buysse, M. De Leenheer, C. Develder, B. Dhoedt, Exploiting relocation to reduce network dimensions of resilient optical grids, in: Proc. 7th Int. Workshop Design of Reliable Commun. Network (DRCN 2009), Washington, DC, USA, 2009, pp. 100–106, doi:10.1109/DRCN.2009.5340020.

[19] C. Develder, B. Mukherjee, B. Dhoedt, P. Demeester, On dimensioning optical grids and the impact of scheduling, Photonic Network Commun. 17 (3) (2009) 255–265, doi:10.1007/s11107-008-0160-z.

[20] K. Christodoulopoulos, E. Varvarigos, C. Develder, M. De Leenheer, B. Dhoedt, Job demand models for optical grid research, in: Proc. 11th Int. IFIP TC6 Conf. on Optical Netw. Design and Modeling (ONDM2007), Athens, Greece, 2007, pp. 127–136.

[21] B. Van Houdt, C. Develder, J.F. Pérez, M. Pickavet, B. Dhoedt, Mean field calculation for optical grid dimensioning, IEEE/OSA J. Opt. Commun. Network 2 (6) (2010) 355–367, doi:10.1364/JOCN.2.000355.

Please cite this article in press as: J. Buysse et al., Providing resiliency for optical grids by exploiting relocation: A dimensioning study based on ILP, Comput. Commun. (2011), doi:10.1016/j.comcom.2010.12.014